How learning shapes the empathic brain

Hein, Grit ; Engelmann, Jan B ; Vollberg, Marius C ; Tobler, Philippe N

Abstract: Deficits in empathy enhance conflicts and human suffering. Thus, it is crucial to understand how empathy can be learned and how learning experiences shape empathy-related processes in the human brain. As a model of empathy deficits, we used the well-established suppression of empathy-related brain responses for the suffering of out-groups and tested whether and how out-group empathy is boosted by a learning intervention. During this intervention, participants received costly help equally often from an out-group member (experimental group) or an in-group member (control group). We show that receiving help from an out-group member elicits a classical learning signal (prediction error) in the anterior insular cortex. This signal in turn predicts a subsequent increase of empathy for a different out-group member (generalization). The enhancement of empathy-related insula responses by the neural prediction error signal was mediated by an establishment of positive emotions toward the out-group member. Finally, we show that surprisingly few positive learning experiences are sufficient to increase empathy. Our results specify the neural and psychological mechanisms through which learning interacts with empathy, and thus provide a neurobiological account for the plasticity of empathic reactions.

DOI: https://doi.org/10.1073/pnas.1514539112

How learning shapes the empathic brain

Grit Hein[1], Jan B. Engelmann[1,2], Marius Vollberg[3], and Philippe N. Tobler[1]

[1] Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, 8006 Zurich, Switzerland

[2] Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, 6525, The Netherlands

[3] Department of Experimental Psychology, Faculty of Brain Sciences, University College London, London, WC1E 6BT, Great Britain

Corresponding author: Grit Hein, Blümlisalpstrasse 10, CH-8006 Zürich

email: grit.hein@econ.uzh.ch

Keywords: empathy, learning, ingroup, fMRI

Abstract

Deficits in empathy enhance conflicts and human suffering. Thus, it is crucial to understand how empathy can be learned, and how learning experiences shape empathy-related processes in the human brain. As a model of empathy deficits, we used the well-established suppression of empathy-related brain responses for the suffering of outgroups, and tested whether and how outgroup empathy is boosted by a learning intervention. During this intervention, participants equally often received costly help, either from an outgroup member (experimental group) or from an ingroup member (control group). We show that receiving help from an outgroup member elicits a classical learning signal (prediction error) in the anterior insular cortex. This signal in turn predicts a subsequent increase of empathy for a different outgroup member (generalization). The enhancement of empathy-related insula responses by the neural prediction error signal was mediated by an establishment of positive emotions towards the outgroup member. Finally, we show that surprisingly few positive learning experiences are sufficient to increase empathy. Our results specify the neural and psychological mechanisms through which learning interacts with empathy, and thus provide a neurobiological account for the plasticity of empathic reactions.

Significance statement

Deficits in empathy for outgroup members are pervasive, with negative societal impact. It is therefore important to ascertain whether empathy towards outgroups can be learned and how learning experiences change empathy-related brain responses. We used a learning intervention during which participants experienced help from a member of their own social group or of a generally depreciated outgroup. Our results show that the intervention successfully increased empathy-related brain responses towards the outgroup. These changes in outgroup empathy were triggered by the learning signal (prediction error) elicited during the first two positive outgroup experiences. Together, our results show that classical learning signals update empathic brain responses and that surprisingly few positive experiences with an outgroup member are sufficient to increase outgroup empathy.

\body Empathy deficits have detrimental social effects (1). When they concern outgroups, these deficits are particularly pervasive (2) and expressed in brain regions that are related to empathy processing (3-9), for example the anterior insular cortex (7-9). Given the multicultural nature of our societies, scholars from various disciplines have aimed to increase empathy for outgroups. They report evidence that positive intergroup contact, for example broadcasted in a radio drama (10) or experienced in an intergroup workshop (11), can increase empathy for outgroup members. However, the mechanisms underlying such changes in empathy are poorly understood, which impedes the development of principled interventions to foster empathy.

Based on the finding that positive intergroup contact is beneficial, one can plausibly assume that increases in empathy towards another person can be achieved via the establishment of positive associations with that person (12). The mechanisms which drive the establishment of positive associations have been heavily studied in the domain of learning theory (13-18). A learning-theoretical framework predicts that the establishment of positive associations towards a person is most efficient, if the actions of that person result in unexpected positive outcomes. This is because unexpected positive outcomes elicit *positive prediction errors* (18-20), that is, a large difference between the learner's prior (low) expectation and the positivity of the actual outcome. Based on this rationale, empathy for a person is learned if the person's actions elicit a prediction error signal (i.e. yields an unexpected positive outcome) that drives the establishment of positive associations. This increased positivity towards the other person, in turn, should raise empathy (12).

These assumptions provide a clear and testable mechanism how learning can increase empathy. However, so far little is known about the interplay between classical learning mechanisms and empathic processes, and how learning experiences shape empathy-related processes in the human brain. Here we use fMRI, combined with formal learning theory and

an intergroup conflict paradigm to investigate whether and how classical learning mechanisms alter the empathy of males towards outgroup members.

Rationale

Our study consisted of three parts, a pre-intervention part, a learning intervention, and a post-intervention part. To investigate the interaction between learning and empathy we exploited an ecologically valid intergroup conflict in our country (Switzerland). During all three parts of the study, the Swiss participants were paired with individuals of Swiss descent (ingroup members) and individuals of Balkan descent (outgroup members). The latter form a large minority in Switzerland whose presence is often portrayed as problematic.

The learning intervention was based on the principles of negative reinforcement, i.e. learning that arises from the absence of a negative outcome. The participant expected to receive painful shocks. However, he knew that one of the other individuals in the scanner room could give up money to save him from pain (Fig. 1). The name of the potential helper was revealed just before the intervention started, and was a typical Balkan name in the experimental, and a typical Swiss name in the control group. Apart from these differences in names, the intervention was identical in both treatment groups.

To measure empathy, we assessed participants' brain responses while they were observing pain in the ingroup and in the outgroup member, which is a well-established procedure for assessing neural activation related to empathy for pain (9, 21). Pain stimulation on the back of the ingroup or outgroup member's hand was indicated by visual cues (see Methods). Importantly, before and after the intervention the ingroup and the outgroup member were represented by different individuals. This setup allowed us to test whether and how the learning intervention affects the neural response to ingroup and outgroup pain and whether its potential impact generalizes to other individuals who were not present during the intervention, but were members of the same respective groups.

Given that expectations concerning outgroup members are typically more negative than those concerning ingroup members (22), receiving help from an outgroup member is an unexpected positive outcome which should elicit a strong positive prediction error. If so, then the participants of the experimental group should arguably use the prediction errors to establish positive associations with their outgroup helper. This increase in outgroup positivity should in turn increase empathy for the suffering of the outgroup member (12), reflected by an increase of activation in empathy-related brain regions such as the AI (23, 24). In contrast, the participants of the control group are likely to expect help from the ingroup member. Thus, even though the control group is saved from pain exactly as often as the experimental group, we predicted less learning and no significant change in empathy in the control condition.

Results

For the analyses of behavioral data we focused on ingroup vs. outgroup differences in impression ratings and emotion ratings. Impression ratings (9, see also SI) served as a manipulation check for the group manipulation and were collected before scanning. Participants had significantly more positive impressions of the ingroup members, as compared to the outgroup members, $F(1,36) = 12.5$, $P = 0.001$, experimental group, $t(1,19) = 2.6$, $P = 0.018$; control group, $t(1,17) = 2.5$, $P = 0.02$. Emotion ratings served to determine whether the learning intervention had established positive outgroup associations, and were collected at the end of each intervention trial. We used a linear regression model to compare learning-related changes in emotion ratings in the experimental versus the control group. To account for potential differences between early and late learning stages (25, 26), we also compared the effects in the first (Trials 1-10) and the second half (Trials 11-20) of the intervention. Emotions towards the outgroup member (experimental group) became more positive than those towards the ingroup member (control group), in particular in the first half of the intervention, treatment (experimental / control group) x trial (Trial 1-20) x half (first / second half), $T = 2.1$, $P = 0.03$; treatment x trial interaction, first half, $T = -1.95$, $P = 0.05$ (Fig. S1),

second half, $T = 1.29$, $P = 0.2$ (Table S1). These results show that in the experimental group, the learning intervention established positive associations towards the outgroup member, with particularly strong effects in the first half.

Next, we tested if the establishment of positive outgroup associations in the experimental group had an impact on empathy-related brain responses after the intervention as compared to before the intervention. Based on previous studies (7-9), we assumed that potential learning effects on outgroup empathy should modulate the neural response in the anterior insula (AI). To test this assumption, we analyzed our data in bilateral anatomical masks of the insular cortex (27), using small-volume family-wise error (SV FWE) correction (see Supplementary tables for whole brain results). Before the intervention, participants' brain responses in the left AI were stronger when they saw the ingroup member as compared to the outgroup member in pain; experimental group, $T = 6.16$, $Z = 4.14$, SV FWE-corrected; control group, $T = 5.71$, $Z = 3.9$, SV FWE-corrected (Table S2). Confirming previous results (7-9), these findings show an outgroup deficit in the AI when comparing observed ingroup to outgroup pain.

To investigate whether the learning intervention counteracted this outgroup deficit, we calculated group (ingroup/outgroup) x time (pre-intervention/post-intervention) interactions for each participant, and compared the average interaction contrasts between the experimental and the control group, using a two-sample *t*-test. Note that this approach is testing a three-way interaction between group (ingroup/outgroup), time (pre-intervention/post-intervention), and treatment (experimental group/control group), while providing information about the direction of the effect at the same time (experimental group greater than control group). The main result was a significant intervention effect in bilateral AI, which was more substantial in the left hemisphere, $T = 4.72$, $Z = 4.13$, SV FWE-corrected (Fig. 2a; Table S3).

In follow-up analyses, we tested for a group (ingroup/outgroup) x time (pre-intervention/post-intervention) interaction, for each (the experimental and the control) group

separately. For the experimental group, we again found activation in the left AI, $T = 4.41$, $Z = 4.15$, SV FWE-corrected. In contrast, in the control group, the interaction revealed no significant activations, even at a relaxed threshold of $P < 0.05$, uncorrected. Thus, the intervention effect in AI cortex is based on pre-to-post increases in AI activation in the experimental, but not in the control group (Fig. 2b). Moreover, AI activation increased primarily for outgroup members in pain. This was evidenced by a significant difference in the outgroup post vs pre contrast between the experimental group and control group in left AI, $T = 4.25$, $Z = 3.8$, SV FWE-corrected. There was no difference between the groups with regard to the ingroup conditions (ingroup post vs. pre), even at $P < 0.05$ uncorrected. Together, these results demonstrate a significant intervention effect in left AI, which reflects an increase in the experimental group participants' neural response to the outgroup member's pain.

Importantly, the neural response in the left AI region, which changed in response to the intervention, correlated with self-reported empathy, that is, the individual ratings on the Empathic Concern Scale (28), which we collected after scanning, $r(38) = 0.39$, $P = 0.015$ (Fig. 2c). This finding confirms that the observed neural effects in this region are related to empathy.

After establishing the success of our intervention, we investigated the mechanisms underlying it. We predicted that positive experiences, i.e., receiving help would elicit positive prediction errors, in particular in the case of an outgroup helper. We computed the individual prediction errors based on a modified version of a reinforcement learning model (19), which allowed us to model learning as a function of individual prior expectations. The impression ratings reported above indicate that participants entered the learning intervention with more positive expectations towards the ingroup as compared to the outgroup member. To capture the individual variability in expectations, our model included each participant's impression ratings for the outgroup members (experimental group) and the ingroup members (control group) as starting predictions for subsequent prediction error estimation (see Methods). This

approach extends classical prediction error models in which learning starts from a starting prediction of zero (the assumption being that participants have neutral prior expectations).

First, we identified brain regions that track individual prediction errors by regressing neural activity elicited by the decisions of the potential helper (outgroup member in the experimental group and ingroup member in the control group) against trial-by-trial estimates of prediction errors. The results revealed activation in the AI (Fig. 3a, red), which was more substantial in the right AI, $T = 4.56$, $Z = 4.01$, SV FWE-corrected (Table S4).

Second, we tested whether the individual prediction error signal in right AI predicts the pre-to-post changes in empathy, that is, the neural intervention effect shown in Fig. 2a. For each individual, we extracted the prediction error-related beta values from right AI and regressed them against the individual magnitude of the intervention effect, as reflected by the group (ingroup/outgroup) x time (pre intervention/post intervention) interaction. The results showed significant activation in left AI, $T = 5.28$, $Z = 4.52$, SV FWE-corrected (Fig. 3b, red, Table S5), which overlapped with the observed intervention effect (Fig. 3b, orange, Fig. 2a for comparison). This result shows that the change in empathy after compared to before the intervention is linked to the magnitude of the learning signal during the intervention.

Third, we assumed that the prediction error signal affects empathy-related brain responses via the establishment of positive associations with the outgroup member. Accordingly, the direct impact of the prediction error signal on the intervention effects should be mediated by the learning-related increase in positivity towards the helper. We identified brain signals that were related to a learning-related increase in positivity by correlating the trial-by-trial emotion ratings with the neural response when the helper's decision was revealed. The results showed significant activation in right AI, $T = 4.7$, $Z = 4.1$, SV FWE-corrected (Fig. 3a, yellow, Table S6), which overlapped with the prediction error signal (Fig. 3a, orange).

Bootstrapped mediation analyses (29) were then conducted to examine whether the neural change in positivity mediated the effect of the neural prediction error signal on the empathy intervention effect (SI). The results revealed a significant indirect path from the prediction error signal, via the neural increase in positivity, to the neural intervention effect, *B* = 3.4, 95% confidence interval (CI) = [0.15 to 7.9]. Furthermore, after controlling for the indirect path, the significant correlation between the prediction error signal and the intervention effect (c) became non-significant (Fig. 3c, comparison between c and c'), reflecting full mediation. Together, these results corroborate the hypothesis that positive prediction errors drive pre-to-post changes in empathy as mediated by the establishment of positive emotions.

In real life, the number of positive interactions with an outgroup member is likely to be small. It is therefore of practical importance to investigate the minimal number of positive outgroup experiences that is necessary to predict increases in outgroup empathy. So far, we have shown that a successful intervention effect is obtained after fifteen positive outgroup experiences, but it would be useful to know whether a smaller number of positive learning experiences with the outgroup is sufficient to increase outgroup empathy. To explore which phase of the learning intervention is most effective, we divided the entire period in quarters of equal length (Trials 1-5, 6-10, 11-15, 16-20), and extracted the respective prediction-error-related activity from right AI (30 for a similar approach). We then tested the predictive relation between each of these four intervention phases on the individual post-minus-pre intervention change in left AI activation (see SI). The results showed that the initial five intervention trials alone accounted for 62.5% of the variance in the intervention effect, which yielded a significant model, $F(1,19) = 11.5$, $P = 0.003$. The prediction error-related activity during the initial five learning trials remained the best predictor even when all the other predictors were added. In fact, none of the other three predictors contributed significantly to

explaining individual variance in the intervention effect (Table 1), Model 2, R^2change = 0; Model 3, R^2change = 0.017; Model 4, R^2change = 0.023.

The initial five intervention trials contained an average of four positive learning experiences (M = 3.8, SD = 0.91). To further specify the number of positive experiences required to predict the individual intervention effects, we extracted the right AI prediction error signal related to the first, second, third, and fourth positive learning experience and assessed their relationship with the individual intervention effect in separate regression analyses. For each regression, the number of negative experiences (i.e., when the participant did not receive help) that preceded the individual number of positive experiences was included as a control variable. The results showed that the individual magnitude of the intervention effect is predicted after only 1 to 2 positive learning experiences with the outgroup member, first positive experience, B = 3.04, T = 2.43, P(FDR-corrected) = 0.053, second positive experience, B = 4.4, T = 2.9, P(FDR-corrected) = 0.038. By contrast, the prediction error signals elicited by three and four positive experiences no longer predicted the intervention effect, three positive experiences, B = 2.6, T = 1.06, P(FDR-corrected) = 0.4, four positive experiences, B = 0.36, T = -0.12, P(FDR-corrected) = 0.9.

Discussion

Our findings show that empathy with an outgroup member can be learned and generalized. This learning is driven by classical prediction errors, whose impact on empathy signals is mediated by an increase in positivity towards the outgroup member. Thus, our study provides a mechanistic account of how positive contacts with the outgroup can counteract empathy deficits (10, 11).

The reductions in outgroup deficits were predictable after surprisingly few (two) positive experiences with an outgroup individual. Although the exact number of experiences should be taken with a grain of salt, it is unlikely that these findings reflect noise, for example induced by a specific constellation of helping and non-helping trials in the beginning of the

intervention. First, an analysis that was based on the full data set (and therefore less prone to noise) independently identified the strongest learning effects in the first five trials (Table 1). Second, a detailed trial-by-trial analysis revealed that the average ratio of helping to non-helping trials within the first two trials was not different from that ratio in later phases of the learning intervention, indicating that volatility does not change in early compared to late phases of the experiment (Supplementary Results). Third, prediction error signals during the first two intervention trials were similar, irrespective of whether participants had experienced no or some non-helping trials (Supplementary Results). Fourth, in the analysis that revealed a predictive relation between the AI-activity during the intervention and subsequent empathy change we controlled for the number of non-helping trials experienced by each subject.

So far, efficient learning based on a very few events has mainly been shown in the domain of punishment. For example, in animals, aversive reactions can be learned based on a single aversive event (31, 32), and humans need only two to three negative experiences (33). Going beyond this work, our findings show that people can learn very efficiently from positive social experiences that prevent them from harm, and that this type of learning has a strong impact on complex internal states, such as empathy for a person in pain. Thus, we provide novel insights into the efficiency of negative reinforcement learning in humans and its potential to serve as an intervention to counteract deficits in outgroup empathy.

Moreover, we found that the learning experience, which was initiated by one representative of the outgroup, resulted in an increase in empathy for another representative of the outgroup who was in pain. The generalization of the learning effect is important, because it shows the robustness of the learning intervention and its potential relevance for society (1). The complex experimental set up (including cover story and confederates) did not allow for repeating the empathy measure for a second time to test for long-term effects. However, it has been shown that negative reinforcement can induce robust and long-lasting learning effects,

for example in therapeutic settings (34). In the light of such results it is conceivable that learning interventions like ours might elicit lasting intervention effects.

Our results reveal how learning underpins the dynamics of empathy. Unexpected positive outcomes resulting from help of another person elicit prediction error signals, that is, signals that resemble the ones known to drive reinforcement learning in the monkey brain (17, 18). Our results indicate that basic learning mechanisms are also used during complex social learning, which is in line with previous studies (13, 35-38). Going further, our findings show how classical learning mechanisms shape other-regarding motivational states such as empathy.

We find that learning about another person and experiencing empathy for a person in pain recruit a common neural structure, namely, the AI cortex (see SI for a discussion of lateralization). Different fields of research have independently accumulated evidence for the important role of the AI in the processing of empathy (23) and the encoding of prediction errors during learning (15, see Table S7, Supplementary Discussion for less significant striatal effects). Our results integrate these two domains. We show that, during learning, the AI is involved in updating predictions about future outcomes as well as in implementing the resulting emotional states. Interestingly, the updated information is used to modulate the empathic reaction to another person. Based on these results, an empathy-learning model would propose that empathy-related processes in AI are altered by a person's individual learning history.

According to the empathy-learning model, empathic responses are altered by any information that elicits prediction errors and thereby results in an update of predictions about others. Thus, it makes the clear prediction that empathy learning should be the stronger, the more positive and unexpected the information revealed about another person. These predictions of the empathy-learning model provide a plausible mechanism for the effect of positive intergroup contact (10, 11), and can inspire new interventions to foster empathy. On

the conceptual level, our results uncover the neural interplay between empathy and learning, and thus provide a neurobiological mechanism for the profound plasticity of empathic reactions, which has been widely documented (39, 40), but so far not explained.

Methods

Participants. Forty healthy men (mean age = 22.7, SE = 0.41) participated in the study. They were randomly assigned to the experimental and the control group with no age difference between the groups, $t(38) = -0.34$, $P = 0.73$. We chose a male instead of a gender-mixed participant group because it allowed us to also choose male confederates and avoided the potential complications of gender-mixed pairing of participants and confederates. Moreover, testing the modulation of empathy in males is more conservative than in females, because males are less likely to simulate others' emotional state on the neural level (41). Two datasets of the control group had to be excluded because of technical problems during fMRI data collection. Participants gave informed consent and the study was approved by the Research Ethics Committee of the Canton of Zurich.

Pre-scanning procedure. We used a well-established priming procedure (42) to induce the ingroup–outgroup manipulation and to activate the relevant stereotype. Details about the pre-scanning procedure and the cover story, are provided as SI.

Scanning procedure. During the *Pre-intervention empathy session,* the participant in the scanner observed the ingroup or the outgroup confederate receive painful stimulation (18 trials each). Each trial started with an arrow cue (500ms), whose color indicated the recipient of the pain (ingroup / outgroup member). After a fixation period (1500ms), a lightning bolt was presented whose color matched the color of the arrow cue (1000ms). Next, the color of the bolt changed to yellow, which indicated the delivery of the painful stimulation to the respective person (1000ms). After a fixation period (1000-3000ms), the next trial was presented. The colors indicating the ingroup and outgroup condition were counter-balanced

across participants. The trials were presented in pseudo-randomized order (no more than two consecutive trials of the same condition). The *Intervention session* consisted of 20 trials, 15 trials in which the participant received help from the other person, and five in which he did not receive help and was thus subjected to pain. Helping and non-helping trials were presented in random order (for details see Fig. 1). The *Post-intervention session* was identical to the pre-intervention session, except that the participant observed the painful stimulation of a new ingroup and a new outgroup member. The participants (and confederates) were informed that they would not meet after the study and had separate visual displays to keep emotion ratings anonymous.

Prediction error model. Prediction errors were computed according to $\delta_t = \alpha\,(\lambda_t - V_t)$, where V_t corresponds to the value V predicted by all stimuli presented in trial t, λ_t corresponds to the value of the outcome in trial t, and α corresponds to the learning rate. The learning rate determines how much weight is given to recent experience as captured by the prediction error. We assumed a learning rate of 0.3, which is most commonly reported in reinforcement learning paradigms (43). We found that the prediction-error-related effects were very similar for other learning rates (0.2 and 0.4). Additional prediction error estimates based on the emotion ratings confirmed the applied learning rate (Supplementary Results). To capture individual variability in prior expectations, V_1 was equal to each participant's average outgroup score (experimental group) or ingroup score (control group) on the impression scale. In contrast to this novel approach, traditional learning theory assumes no prior expectations ($V_1 = 0$), which is unlikely in our social setting. The boundary outcome values were set according to the maximum (54 points) and minimum (6 points) score of the impression scale. Accordingly, $\lambda_t = 54$ in helping trials and $\lambda_t = 6$ in non-helping trials.

Imaging analyses. We conducted standard pre-processing, first and second-level analyses (see SI). Second-level results were corrected for multiple comparisons by using family-wise error correction (FWE) within bilateral anatomical masks of the entire insular cortex, as defined by

the AAL atlas (27). For data extraction, we used the entire cluster of the respective activations. The extracted beta values reflect the average activation of all voxels within the cluster. Details about the multiple regression analyses and the mediation analysis are provided as SI.

Acknowledgments

This study was supported by the Swiss National Science Foundation (Grants PP00P1_128574, PP00P1_150739, and CRSII3_141965). We also acknowledge the Neuroscience Center Zurich. We thank Jessica Gomes for help with data collection, Karl Treiber for scanning support, Chris Burke for valuable input on the data analyses, and Alexander Soutschek, Björn Lindström, and Tamara Herz for helpful comments on the manuscript.

References

1. Baron-Cohen S (2011) *Zero Degrees of Empathy: A New Theory of Human Cruelty* (Penguin, UK).
2. Cikara M, Bruneau EG, Saxe RR (2011) Us and them: Intergroup failures of empathy. *Curr Dir Psychol Sci* 20(3):149-153.
3. Cikara M, Botvinick MM, Fiske ST (2011) Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychol Sci* 22(3):306-313.
4. Cikara M, Fiske ST (2011) Bounded empathy: Neural responses to outgroup targets'(mis) fortunes. *J Cogn Neurosci* 23(12):3791-3803.
5. Cheon BK, et al. (2011) Cultural influences on neural basis of intergroup empathy. *Neuroimage* 57(2):642-650.
6. Bruneau EG, Dufour N, Saxe R (2012) Social cognition in members of conflict groups: Behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. *Philos Trans R Soc Lond B: Biol Sci* 367(1589):717-730.
7. Azevedo RT, et al. (2013) Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Hum Brain Mapp* 34(12):3168-3181.
8. Contreras-Huerta LS, Baker KS, Reynolds KJ, Batalha L, Cunnington R (2013) Racial bias in neural empathic responses to pain. *PLoS One* 8(12):e84001.
9. Hein G, Silani G, Preuschoff K, Batson CD, Singer T (2010) Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron* 68(1):149-160.
10. Paluck EL (2009) Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *J Pers Soc Psychol* 96(3):574-587.
11. Malhotra D, Liyanage S (2005) Long-term effects of peace workshops in protracted conflicts. *Journal of Conflict Resolution* 49(6):908-924.

12. Batson CD, Eklund JH, Chermok VL, Hoyt JL, Ortiz BG (2007) An additional antecedent of empathic concern: Valuing the welfare of the person in need. *J Pers Soc Psychol* 93(1):65-74.
13. Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. *Nature* 456(7219):245-249.
14. Bhanji JP, Delgado MR (2014) The social brain and reward: Social information processing in the human striatum. *Wiley Interdiscip Rev Cogn Sci* 5(1):61-73.
15. Bossaerts P (2010) Risk and risk prediction error signals in anterior insula. *Brain Struct Funct* 214(5-6):645-653.
16. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442(7106):1042-1045.
17. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593-1599.
18. Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307(5715):1642-1645.
19. Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, eds Black AH, Prokasy, William F(Appleton-Century-Crofts, New York), pp 64-99.
20. Steinberg EE*,* et al. (2013) A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* 16(7):966-973.
21. Singer T*,* et al. (2004) Empathy for pain involves the affective but not sensory components of pain. *Science* 303(5661):1157-1162.
22. Dovidio JF, Gaertner SL (1993) Stereotypes and evaluative intergroup bias. *Affect, Cognition, and Stereotyping: Interactive Processes in Group Perception*, eds Mackie DM, Hamilton, David L(Academic Press, San Diego), pp 167-193.
23. Lamm C, Decety J, Singer T (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage.* 54(3):2492-2502.
24. Gu X*,* et al. (2010) Functional dissociation of the frontoinsular and anterior cingulate cortices in empathy for pain. *J Neurosci* 30(10):3739-3744.
25. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38(2):329-337.
26. Tobler PN, Fletcher PC, Bullmore ET, Schultz W (2007) Learning-related human brain activations reflecting individual finances. *Neuron* 54(1):167-175.
27. Tzourio-Mazoyer N*,* et al. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15(1):273-289.
28. Davis MH (1983) Measuring individual differences in empathy: Evidence for a multidimensional approach. *J Pers Soc Psychol* 44(1):113-126.
29. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 40(3):879-891.
30. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66(4):585-595.
31. Keith-Lucas T, Guttman N (1975) Robust single-trial delayed backward conditioning. *J Comp Physiol Psychol* 88(1):468-476.

32. Mahoney WJ, Ayres JJ (1976) One-trial simultaneous and backward fear conditioning as reflected in conditioned suppression of licking in rats. *Anim Learn Behav* 4(4):357-362.
33. Steinberg C, Bröckelmann A-K, Rehbein M, Dobel C, Junghöfer M (2013) Rapid and highly resolving associative affective learning: Convergent electro-and magnetoencephalographic evidence from vision and audition. *Biol Psychol* 92(3):526-540.
34. Iwata BA, Pace GM, Kalsher MJ, Cowdery GE, Cataldo MF (1990) Experimental analysis and extinction of self-injurious escape behavior. *J Appl Behav Anal* 23(1):11-27.
35. Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. *Pro Natl Acad Sci* 107(32):14431-14436.
36. Seid-Fatemi A, Tobler PN (2015) Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. *Soc Cogn Affect Neurosci* 10(5):735-743.
37. Suzuki S, et al. (2012) Learning to simulate others' decisions. *Neuron* 74(6):1125-1137.38.
38. Fareri DS, Chang LJ, Delgado MR (2015). Computational substrates of social value in interpersonal collaboration. *J Neurosci*. 35(21):8170-8180.
39. Hein G, Singer T (2008) I feel how you feel but not always: the empathic brain and its modulation. *Curr. Opin. Neurobiol.* 18(2):153-158.
40. Zaki J (2014) Empathy: A motivated account. *Psych. Bull.* 140(6):1608-1647.
41. Schulte-Rüther M, Markowitsch HJ, Shah NJ, Fink GR, & Piefke M (2008) Gender differences in brain networks supporting empathy. *Neuroimage* 42(1):393-403.
42 Dijksterhuis A, Van Knippenberg A (1998) The relation between perception and behavior, or how to win a game of trivial pursuit. *J. Pers. Soc. Psychol.* 74(4):12.
43 Gershman SJ (2015) Do learning rates adapt to the distribution of rewards? *Psychon Bull Rev*. 22(5): 1320-1327.

Figure Legends

Figure 1. Example trial of the learning intervention. The arrow cue indicated painful stimulation for the participant. Next, the options for the potential helper (outgroup member in the experimental group, ingroup member in the control group) were shown. By choosing the crossed-out lightning bolt symbol, the potential helper indicated his decision to give up five Swiss francs to cancel delivery of pain stimulation to the participant. By choosing the intact lightning bolt symbol, he indicated his decision to keep the money, which led to a painful shock for the participant at the end of the trial. The potential helper's decision was highlighted with a yellow square. The participant rated how he felt about the potential helper on an emotion rating scale. In this example, the potential helper's decision canceled delivery of the pain stimulation, indicated by a crossed-out lightning bolt at the end of the trial. Otherwise,

the intact lightning bolt was presented, and the participant received a painful shock. To allow for successful positive conditioning, the participant was saved from pain in 75% of all trials (15 out of 20 trials).

Figure 2. Impact of the intervention on neural responses during observation of pain in the ingroup and outgroup member in the experimental and control group. **a)** Significant activation in bilateral anterior insular cortex (AI), indicating stronger intervention-related effects in the experimental group as compared to the control group (Table S3). **b)** Average parameter estimates for the contrast between observing ingroup pain versus outgroup pain in left AI. As a result of the intervention, empathy responses to outgroup pain compared to ingroup pain were elevated in the experimental group, but remained biased for ingroup pain in the control group. **c)** Positive correlation between the individual ratings on the Empathic Concern Scale of the Interpersonal Reactivity Index (28) and the neural response in left AI to ingroup pain prior to the intervention in the experimental group (black circles) and the control group (gray triangles). We chose the neural response in the ingroup condition prior to the intervention because this measure is most likely to reflect participants' trait empathy (i.e., their tendency to empathize irrespective of the effects of the outgroup manipulation or the intervention). Note that the extracted data are independent of the statistical analysis that defined the extraction region in left AI (i.e., the group x time x treatment interaction). For additional correlation analyses see Figure S2. Error bars represent standard errors. The imaging results are displayed at FWE corrected < 0.05 (SV in bilateral anatomical masks of the insular cortex).

Figure 3. Neural responses correlating with trial-wise prediction errors and emotion ratings during the intervention and their impact on the neural intervention effect, that is, the individual (ingroup/outgroup) x time (pre-intervention/post-intervention) interaction. **a)**

Neural response in right AI reflects prediction errors (red) and emotion ratings (yellow). The overlap (orange) indicates that the same region in right AI encodes prediction errors and increasing positive emotions (Tables S4 and S6). **b)** Neural intervention effect in left AI predicted by the individual prediction errors (red). The predicted intervention effect overlaps with the actual intervention effect (orange, see Fig. 2a for comparison; Table S5). **c)** Results of the mediation analysis. The indirect path from the prediction error signal to the neural increase in positive emotions (a) to the intervention effect (b) was significant. The direct impact of the prediction error on the intervention effect (c) became non-significant after controlling for the indirect path (c'). This indicates that the effect of the prediction errors on pre-to-post changes in empathy is fully mediated by changes in emotions towards the helper. Numbers indicate beta coefficients, numbers in parentheses standard errors. *$P < 0.05$, ** $P < 0.01$. The imaging results are displayed at FWE corrected < 0.05 (SV in bilateral anatomical masks of the insular cortex).